LAMP-TR-101
CS-TR-4456
UMIACS-TR-2003-26

February 2003

CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval

Kareem Darwish, Douglas W. Oard

Language and Media Processing Labratory
Institiue for Advanced Computer Studies
College Park, MD 20742

Abstract

The focus of the experiments reported in this paper was techniques for combining evidence for cross-language retrieval, searching Arabic documents using English queries. Evidence from multiple sources of translation knowledge was combined to estimate translation probabilities, and four techniques for estimating query-language term weights from document-language evidence were tried. A new technique that exploits translation probability information was found to outperform a comparable technique in which that information was not used. Comparative results for three variants of Arabic light stemming are also presented. A simple variant of an existing stemming algorithm was found to result in significantly better retrieval effectiveness.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

Report Documentation Page	Form Approved OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE **CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Institute for Advanced Computer Studies,Language and Media Processing Laboratory,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**9**	

CLIR Experiments at Maryland for TREC-2002:
Evidence combination for Arabic-English retrieval

Kareem Darwish and Douglas W. Oard[1]
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
{kareem,oard}@glue.umd.edu

Abstract

The focus of the experiments reported in this paper was techniques for combining evidence for cross-language retrieval, searching Arabic documents using English queries. Evidence from multiple sources of translation knowledge was combined to estimate translation probabilities, and four techniques for estimating query-language term weights from document-language evidence were tried. A new technique that exploits translation probability information was found to outperform a comparable technique in which that information was not used. Comparative results for three variants of Arabic "light" stemming are also presented. A simple variant of an existing stemming algorithm was found to result in significantly better retrieval effectiveness.

1 Introduction

Statistical techniques for processing natural language are imperfect, but reliance on multiple sources of evidence can help to mitigate the limitations of any one technique. In this paper, we leverage the combination of evidence in two ways: (1) to estimate translation probabilities from multiple sources of translation knowledge, and (2) to estimate weights for a query term weights based on the statistics of multiple document-language terms. Five translation resources of three types (machine translation lexicons, a printed bilingual dictionary that had been manually rekeyed, and translation probabilities derived from parallel corpora) were used as a basis for estimating translation probabilities. Four ways to estimate query term weights were tried (translation-based indexing, Pirkola's structured query method, and two newly developed variants of structured queries). The main task in the TREC-2002 CLIR track was retrieval of Arabic documents using English queries, so we also made some refinements to our Arabic text processing techniques. Most importantly, we compared three approaches to "light" (i.e., one-level rule-based truncation) stemming. Light stemming was found to perform well for retrieval of character-coded Arabic text in TREC-2001 [1, 2], so we limited our experiments to that technique. We begin with a description of the techniques that we tried in the next section. Section 3 then presents our results, and section 4 concludes the paper.

2 Methodology

In this section, the approaches mentioned in the introduction will be described in detail.

2.1 Translation Resources

Using multiple sources of evidence to guide the translation process can help in two important ways: (1) No single source provides a comprehensive set of translations, and (2) No single source provides an accurate indication of translation probabilities. Drawing translation knowledge from diverse sources therefore offers

[1] College of Information Studies and Institute for Advance Computer Studies.

an interesting potential to improve CLIR effectiveness. We had five translation resources of three types available:

- Two bilingual term lists that were constructed using two Web-based machine translation systems (Tarjim and Al-Misbar [3][4]). In each case, we submitted sets of isolated English words found in a 200 MB collection of Los Angeles Times news stories for translation from English into Arabic [5]. Each system returned at most one translation for each submitted word. Together, the two term lists covered about 15% of the unique Arabic stems in the AFP collection (measured by using light stemming on both the term list and the collection).
- The Salmone Arabic-to-English dictionary, which was made available for use in the TREC-CLIR track by Tufts University. No translation preference information is provided in this dictionary, but it does include rich markup describing morphology and part-of-speech information. We preprocessed the dictionary without using any of that additional information, thereby creating a bilingual term list. The coverage of the resulting term list, measured in the same way, was about 7% of the unique Arabic stems in the AFP collection.
- Two translation probability tables, one for English-to-Arabic and one for Arabic-to-English. These tables were constructed from tables provided by BBN, which were in turn constructed from a large collection of aligned English and Arabic United Nations documents using the Giza++ implementation of IBM's model 1 statistical machine translation design. The coverage of the of the Arabic-to-English table, measured in the same way, was 29% of the unique Arabic stems in the AFP collection.

We combined the evidence from these sources in the following manner:

1. We selected a direction (English-to-Arabic or Arabic-to-English), and then inverted any resources that were originally provided in the other direction. For the translation probability table, we retained the probabilities for each translation pair and then renormalized the inverted table so that the values of the "probabilities" for each source-language term summed to one. This process likely introduced some error, since probabilities for rare events may not have been accurately estimated.
2. A uniform distribution was used to assign probabilities to the translations obtained from machine translation systems and the Salmone dictionary. Tarjim and Al-Misbar returned at most one translation for an English word, but two English words might share a common translation. When n alternatives were known from a single source, each was assigned a probability of $1/n$.
3. For translation from English to Arabic, the resulting translation probabilities were then combined across sources by summing the probabilities for a given Arabic translation across the sources in which it appeared and then dividing by the number of sources in which the English term had appeared. For example, if Tarjim, Al-Misbar and Salmone contained the English term, with Tarjim containing some specific translation with probability 1.0, Al-Misbar lacking that translation (i.e., assigning it a probability of 0.0), and Salmone assigning it a probability of 0.5 (because two translations were known), then the resulting combined probability would be 1/3 + 0/3 + 0.5/3 = 0.5. Translation probabilities for Arabic-to-English were computed in the same manner, but with the role of each language reversed.

The resulting translation resource contained what appeared to us to be reasonable estimates of translation probabilities, and covered 36% of the unique Arabic stems in the AFP collection.

2.2 Stemming

Three Arabic light stemmers were tested. The first was Al-Stem, the stemmer provided by the organizers for the standard resource run, which was developed by the author at Maryland and modified by Leah Larkey from University of Massachusetts (U Mass). The second was a light stemmer described in Larkey's SIGIR 2002 paper, which we will refer to as the U Mass stemmer [2]. The third was a modified version of the U Mass stemmer in which two additional prefixes identified through failure analysis using the TREC-2001 topics, were removed. Table 2 lists the prefixes and suffixes removed by each stemmer. Of the three, Al-Stem was the most aggressive stemmer.

Stemmer	Prefixes	Suffixes
Al-Stem	wAl, fAl, bAl, bt, yt, lt, mt, wt, st, nt, bm, lm, wm, km, fm, Al, ll, wy, ly, sy, fy, wA, fA, lA, and bA. (والـ، فالـ، بالـ، بتـ، يتـ، لتـ، متـ، وتـ، ستـ، نتـ، بمـ، لمـ، ومـ، كمـ، فمـ، الـ، للـ، وي، لي، في، وا، فا، لا، با)	At, wA, wn, wh, An, ty, th, tm, km, hm, hn, hA, yp, tk, nA, yn, yh, p, h, y, A. (ـات، وا، ون، وه، ـان، تي، ته، تم، كم، هم، هن، ها، ية، تك، نا، ين، يه، ـة، ـه، ي، ا)
Umass Stemmer	Al, wAl, bAl, kAl, fAl, and w (الـ، والـ، بالـ، كالـ، فالـ، و)	hA, An, At, wn, yn, yh, yp, p, h, and y (ـي، ـه، ـة، ية، يه، ين، ون، ـات، ـان، ها)
Modified UMass Stemmer	-- Identical to U Mass, plus ll, and wll (للـ، وللـ)	-- Identical to U Mass --

Table 2: Prefixes and suffixes removed by each light stemmer.

2.3 Combination of Evidence for Alternate Translations

One of the key challenges in CLIR is what to do when more than one possible translation is known. One principled solution to this problem is Pirkola's structured query method [6]. Pirkola used InQuery's synonym operator to estimate the weight of each query term based on document language evidence as follows:

$$TF_j(Q_i) = \sum_{\{k|D_k \in T(Q_i)\}} TF_j(D_k) \qquad (1)$$

$$DF(Q_i) = \left| \bigcup_{\{k|D_k \in T(Q_i)\}} \{d \mid D_k \in d\} \right| \qquad (2)$$

Where Q_i is a query term, D_k is a document term, $TF_j(Q_i)$ is the term frequency of Q_i in document j, $DF(Q_i)$ is the number of documents that contain Q_i, d is a document, and $T_j(D_k)$ is the set of known translations for the term D_k. The key insight here is that translation is viewed as a process akin to stemming, in which several document-language terms might be treated as if it were the same query language term (in stemming, many morphological variants are treated as if they were the same stem). One limitation of Pirkola's structured query method is that it makes no use of translation probabilities—every possible translation is treated as being equally likely. Consider a case in which one query term has two translations, one of which is highly probable, but which typically has a low term frequency when it appears, while the other is improbable but typically has high term frequencies. Because the improbable translation is so common, documents containing that translation would likely be retrieved ahead of documents that contain the more likely translation. This seems undesirable. Moreover, consider a second case in which an extremely improbable translation appears in many documents, with an almost certain translation appears in only a few. Intuitively, it would seem that we should ignore the extremely improbably translation and assign this term a low document frequency (and thus give it added importance in any retrieval system that relies on inverse document frequency as a measure of term importance). But Pirkola's structured query method does just the opposite.

In this paper, we propose two ways of incorporating translation probability information. The simplest approach is to retain the same formulae, but to suppress the contribution of unlikely translations. We implemented this by starting with the most likely translation and adding additional translations in order of decreasing probability until the cumulative probability of the selected translations reached a preset threshold that was determined through experimentation using the TREC-2001 CLIR collection. As an alternative, we also explored three ways of incorporating translation probabilities directly into the formulae:

1. The weighted DF method (WDF): In this method, translation probability estimates are used in the calculation of document frequency, but not term frequency. Formally, $TF(Q_i)$ is computed as in equation (1), and $DF(Q_i)$ is computed as follows:

$$DF(Q_i) = \sum_{\{k|D_k \in T(Q_i)\}} \left[DF_j(D_k) \times wt(D_k) \right] \qquad (3)$$

Where Q_i is a query term, D_k is a document term, $DF(Q_i)$ is the number of documents that contain Q_i, $wt(D_k)$ is the probability estimate for D_k, and $T_j(D_k)$ is the set of translations for the term D_k. This method addresses the case in which an improbable translation has a high document frequency. Note that the union operator has been replaced by the sum operator, so equations (2) and (3) will not necessarily produce the same value, even in the case of equiprobable translations.

2. The weighted TF method (WTF): In this method, translation probability estimates are used in the calculation of term frequency, but not document frequency. Formally, $DF(Q_i)$ is computed as in equation (2), and $TF(Q_i)$ is computed as follows:

$$TF_j(Q_i) = \sum_{\{k|D_k \in T(Q_i)\}} \left[TF_j(D_k) \times wt(D_k) \right] \qquad (4)$$

   Where Q_i is a query term, D_k is a document term, $TF_j(Q_i)$ is the term frequency of Q_i in document j, and $T_j(D_k)$ is the set of different translations for the term D_k. This method addresses the case in which an improbable translation typically has high term frequencies.
3. The weighted TF/DF method (WTF/DF): In this method, translation probability estimates are used in the calculation of both term frequency and document frequency. Equation (3) is used to compute $DF(Q_i)$, with equation (4) used to compute $TF_j(Q_i)$. This addresses both potential concerns.

The same approach to preselection of the most likely translation probabilities that we used with Pirkola's structured queries can also be used with any of these methods.

2.4 Balanced Translation-Based Indexing

The formulae for Pirkola's method do not depend on any information that can only be computed at query time, so an indexing time implementation is possible. Oard and Ertunc built such an implementation, which they called translation-based indexing. That approach is equivalent to unbalanced document translation, in which every English translation for each Arabic term is indexed once. Unbalanced translation is, however, known to underemphasize the contribution of highly specific terms (which typically have very few possible translations) and to overemphasize the contribution of common terms (which often have many possible translations). The obvious alternative is to balance the translation-based indexing process. To achieve this, we replaced each Arabic term with the five most likely English translations. For terms with fewer than 5 known translations, the most probable translations were replicated in a round-robin fashion until a total of five was reached. We chose the value 5 based on post-hoc experiments with the 25 TREC 2001 topics (we tried 1, 3, 5, 7, and 10).

2.5 Document and Query Expansion

Any individual document will likely include only a fraction of the words that might be used to describe a topic, so some form of expansion to the representation of a document might help retrieval. We therefore prepared a contrastive condition in which the representation of each document was expanded as follows:

1. We identified the 20 most descriptive terms in each document, by dividing the frequency with which each term appeared in the document by the number of documents in which that term was found.
2. We then formed a query with one instance of each of those 20 terms and used that query as a basis for ranking the documents in the AFP collection using the InQuery text retrieval system from the University of Massachusetts (the document being expanded was often in this set). We used the Al-Stem stemmer to normalize the representation of Arabic terms in InQuery, both for query and for document processing.
3. We combined the 10 top-ranked documents into a single mega-document and then chose the 20 most descriptive terms in that mega-document, using the same measure of term importance as above. The resulting set of 20 terms was then added to the representation of document that was being expanded.

One can imagine many variants of this approach, including alternative parameter settings, alternative term importance measures, and alternative ways of combining evidence from the top-ranked documents. Because document expansion was not the principal focus of our experiments, we tried only this one implementation for TREC-2002.

Because queries are relatively brief, they are even more likely to contain only a small subset of the words that might be used to express the concepts that are important for the topic that they represent. We therefore performed pre-translation query expansion using a similar blind relevance feedback process. We used Associated Press articles from 1994-1998 from the North American News Text Corpus (Supplement) and the orld Stream English Collection from the Linguistic Data Consortium for this purpose Because InQuery contains built-in provisions for query expansion, we used InQuery (with the kstem English stemmer) to search these collections. We configured InQuery to choose the most discriminating terms from the top 10 returned documents for each query.

2.6 Implementation Details

To ease work in Arabic, Arabic letters were transliterated to Roman letters. Table 1 shows the mappings between the Arabic letters and their transliterated representations. In preprocessing the text, all diacritics and kashidas were removed, and letter normalization was employed to normalize the letters ya (ي) and alef maqsoura (ى) to ya (ي) and all the variants of alef (ا) and hamza (ء), namely alef (ا), alef hamza (إ ،أ), alef maad (آ), hamza (ء), waw hamza (ؤ), and ya hamza (ئ), to alef (ا).

أ، ا، آ، إ	A	ئ، ؤ، ء	A	ب	b
ت	t	ث	v	ج	j
ح	H	خ	x	د	d
ذ	O	ر	r	ز	z
س	s	ش	P	ص	S
ض	D	ط	T	ظ	Z
ع	E	غ	g	ف	f
ق	q	ك	k	ل	l
م	m	ن	n	ه	h
و	w	ي، ى	y	ة	p

Table 1: English transliteration of Arabic characters

As for a stop-word list, we used the list that is distributed with Sebawai, which includes 130 particles and pronouns [7]. Sebawai is a publicly available Arabic morphological analyzer. We used InQuery for our official monolingual runs and for one of our official cross-language runs. For the remaining monolingual runs and for our post-hoc experiments, we used PSE (an IR system developed at Maryland that uses OKAPI BM-25 weights) because we were not able to perform the necessary changes to the term weight computation using InQuery. Our monolingual and cross-language results should therefore not be considered to be strictly comparable.

3 Results

We submitted five official runs, and we have used the relevance judgments provided by NIST to perform an additional 31 post-hoc runs. In this section we present those results along with a preliminary discussion that we intend to extend in our final paper.

3.1 Arabic Monolingual Runs

We submitted one official automatic monolingual run, one official manual monolingual run, and we performed three post-hoc automatic monolingual runs. In the official automatic run, queries were formed using every word in the title and description fields of the topic description (stopwords were removed, but no automatic stop structure removal was performed). For the manual run, the title, description, and narrative fields were used, additional query terms were manually added, and stop structure and information about what would cause a document to be judged as not relevant was removed manually. The principal goal of the manual run was to enrich the relevance pools. In both cases, Al-stem was used to stem the documents and the queries, and document expansion was used as described above. Table 3 shows the results (results on TREC-2001 data are post-hoc).

The three post-hoc runs were done using PSE rather than InQuery, with a goal of comparing three variants of light stemming. Document expansion was not used in those experiments. Again, Table 3 shows the results. The Al-stem and the U Mass stemmers were found to be statistically indistinguishable by a paired two-tailed t-test (for $p<0.05$). The modified U Mass stemmer did achieve mean average precision results that were statistically better than the two other stemmers over the full 75-topic set, but the advantage over the unmodified U Mass stemmer was not seen when the 50 TREC 2002 topics used alone. Because the suitability of the TREC-2001 collection for post-hoc experiments is open to some question, we conclude that no reliable differences between the three stemmers were detected by our experiments.

Run	Mean Average Precision	
Official Runs (with document expansion)	TREC 2002 Topics (50 topics)	TREC 2001 & 2002 Topics (75 topics)
Automatic Monolingual	0.289	0.314
Manual Monolingual	0.302	0.322
Post-Hoc Runs (no document expansion)	TREC 2002 Topics (50 topics)	TREC 2001 & 2002 Topics (75 topics)
Al-Stem stemmer	0.286	0.316
U Mass stemmer	0.295	0.321
Modified U Mass stemmer	0.301	0.331

Table 3: The Arabic Monolingual Results

3.2 English-Arabic CLIR Runs

We performed three official automatic CLIR runs and 29 post-hoc automatic CLIR runs. In each case, we formed title+description queries in the same manner as for the automatic monolingual run. Pre-translation query expansion was done using blind relevance feedback on Associated Press articles from 1994-1998 using InQuery. The articles were part of the North American News Text Corpus (Supplement) and AP World Stream English Collection from the Linguistic Data Consortium [8]. For the official CLIR runs we tried these following configurations:

- Pirkola's Method. We used InQuery, pre-translation query expansion, document expansion, and the thresholded version of Pirkola's method (with a threshold of 0.3, established using post-hoc experiments with the TREC 2001 topics).
- Balanced Translation-Based Indexing (TBI). We used InQuery, document expansion, and balanced translation-based indexing.
- Weighted TF. We used PSE, pre-translation query expansion, document expansion, and the Weighted TF method (with a threshold of 0.35, again tuned using the TREC-2001 topics).

For the post-hoc experiments, we used PSE, pre-translation query expansion, one of four methods (Pirkola's method, Weighted TF, Weighted DF, or Weighted TF/DF), and a probability threshold that was

varied between 0.1 and 0.7 in increments of 0.1. Post-hoc CLIR results are reported on all 75 topics from TREC 2001 and TREC 2002. We also conducted one additional post-hoc experiment in which we combined the results from our official Weighted TF and translation-based indexing runs in a manner similar to that suggested by [9]. For that experiment, we renormalized the PSE score for the Weighted TF run to be between 0.4 and 0.5 (a range similar to that seen in InQuery), averaged that normalized score with the translation-based indexing score from PSE, and resorted the ranked list based on that averaged score. The result was statistically better than the Weighted TF run, but not statistically distinguishable from the translation-based indexing run. Table 4 shows the results for the official runs and for the one post-hoc experiment that was based on the two official runs. Table 5 and Figure 1 together show the other post-hoc results.

Run	Mean Average Precision	
Official Runs (Expanded documents)	TREC 2002 Topics (50 topics)	TREC 2001 & 2002 Topics (75 topics)
Pirkola's Method	0.202	0.252
Balanced TBI	0.274	0.279
Weighted TF	0.247	0.279
Balanced TBI + Weighted TF	0.289	0.307

Table 4: CLIR results for official runs (expanded documents).

	Cumulative Probability Threshold						
Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Pirkola's Method	0.215	0.212	0.230*	0.230	0.207	0.186	0.134
Weighted TF	0.220	0.226	0.241	0.247*	0.241	0.230	0.192
Weighted DF	0.211	0.196	0.198	0.180	0.147	0.126	0.091
Weighted TF/DF	0.220	0.222	0.235	0.234	0.236	0.240	0.245

Table 5: Post-hoc CLIR results (Mean Average Precision, 75 topics, no document expansion.

The Weighted TF method and translation-based indexing were both statistically significantly better than Pirkola's method (on the 75 topic set), but statistically indistinguishable from each other. The retrieval effectiveness of Weighted TF/DF was close to the best at every cumulative probability threshold, and it was the only one of the four techniques that we tried that did not exhibit a decrease in effectiveness at high thresholds. It therefore seems to be a good candidate for further study, and an appropriate choice if a method

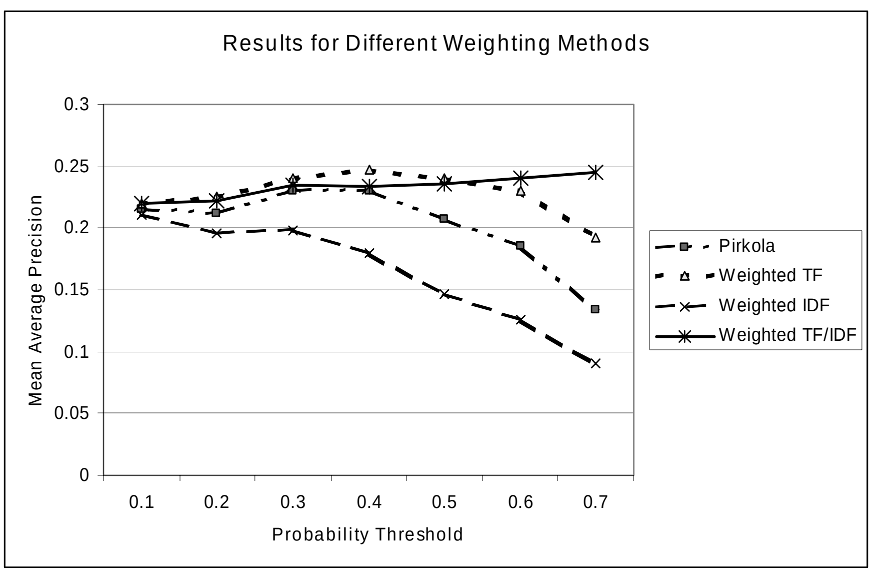

Figure 1: Sensitivity to cumulative probability threshold.

must be selected without access to a representative collection on which to tune the probability threshold. We do not yet know whether document expansion was helpful because the runs with and without document expansion were done using different retrieval systems.

4 Conclusion

We have presented two basic ideas for using a combination of evidence to improve cross-language retrieval effectiveness, demonstrated an ability to produce useful translation probability estimates from multiple sources, and extended Pirkola's structured query method in ways that exploit that information. Our results suggest that further work on the Weighted TF/DF method would be well justified, and that further work on document expansion is needed before the utility of that technique in this context can be judged. We look forward to the discussions that we will have at the conference, and to the opportunity to continue out exploration of these questions through additional post-hoc experiments and analysis.

Acknowledgments

This work was supported in part by DARPA Cooperative Agreement N660010028910.

References

[1] Aljlayl, M., S. Beitzel, E. Jensen, A. Chowdhury, D. Holmes, M. Lee, D. Grossman, and O. Frieder, "IIT at TREC-10," TREC: 265-274, 2001.

[2] Larkey, L., L. Ballesteros, and M. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis," SIGIR 2002: 275-282, 2002.

[3] tarjim.ajeeb.com, Sakhr Technologies, Cairo, Egypt www.sakhr.com

[4] www.almisbar.com, ATA Software Technology Limited, North Brentford Middlesex, UK

[5] NIST, Text Research Collection Volume 5, April 1997.

[6] Oard, W. and F. Ertunc: Translation-Based Indexing for Cross-Language Retrieval. ECIR 2002: 324-333, 2002.

[7] Darwish, K., "Building a Shallow Morphological Analyzer in One Day," ACL Workshop on Computational Approaches to Semitic Languages: 47-54, 2002.

[8] MacIntyre, Robert, "North American News Text Supplement", LDC98T30, LDC, 1998.

[9] McCarley, S., "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval," ACL, 1999.